JOURNAL OF MULTIVARIATE ANALYSIS 19, 273–279 (1986)

# Central Limit Theorem for Perturbed Empirical Distribution Functions Evaluated at a Random Point

MADAN L. PURI*

*Indiana University*

AND

STEFAN S. RALESCU

*Queens College, City University of New York*

*Communicated by M. Rosenblatt*

Let $\hat{F}_n$ be an estimator obtained by integrating a kernel type density estimator based on a random sample of size $n$ from a (smooth) distribution function $F$. Sufficient conditions are given for the central limit theorem to hold for the target statistic $\hat{F}_n(U_n)$ where $\{U_n\}$ is a sequence of $U$-statistics. © 1986 Academic Press, Inc.

## 1. INTRODUCTION

Given a sequence $X_i$, $i \geqslant 1$, of iidrv's (independent and identically distributed random variables) with cdf (cumulative distribution function) $F$, the natural estimator of $F$ based on the sample $X_1, \ldots, X_n$ is the empirical df $F_n$ defined by

$$F_n(x) = n^{-1} \sum_{i=1}^{n} u(x - X_i), \qquad x \in \mathbb{R}, \tag{1.1}$$

where $u(t) = 1$ if $t \geqslant 0$ and $= 0$ otherwise. Although $F_n$ is in a sense already quite smooth, it does not take fully into account the smoothness of $F$ (i.e., the existence of a density $f$). In fact, if $F$ is a continuous cdf, it seems

Received November 30, 1983; revised July 25, 1984

AMS 1980 subject classifications: primary 62E20; 52G30

Key words and phrases: perturbed empirical distribution functions, U-statistics, central limit theorem.

* Research supported by the National Foundation Grant MCS8301409



0047-259X/86 $3.00

reasonable to consider continuous estimators of $F$ which are better adapted to this situation. Thus, in relatively general situations, an estimator of the form

$$\hat{F}_n(x) = n^{-1} \sum_{i=1}^{n} G_n(x - X_i), \qquad x \in \mathbb{R}, \tag{1.2}$$

where $\{G_n\}$ is a sequence of continous cdf's, suggests itself. Furthermore, if one expects this estimator to perform as satisfactory as $F_n$, it seems natural to require that the sequence $\{G_n\}$ be *not* "too far" from the degenerate df $u$, e.g., by requiring that $G_n$ converge weakly to $u$ (written $G_n \to^w u$). Such estimators arise quite naturally as integrals of density estimators of the kernel type. Rosenblatt [7] and Parzen [5] suggested the density kernel type estimator

$$\hat{f}_n(x) = (n\alpha_n)^{-1} \sum_{i=1}^{n} g((x - X_i)/\alpha_n), \qquad x \in \mathbb{R}, \tag{1.3}$$

where $\alpha_n > 0$ and $g$ satisfies $g \geqslant 0$ and $\int_{-\infty}^{\infty} g(t)\, dt = 1$. The interest generated by such estimators is due primarily to their simple structure as averages over the independent elements of the $n$th row of a double array of rv's. We refer to Scott *et al.* [8] and Wertz [12] for a general review in this area. Using the density estimator $\hat{f}_n$, define

$$\hat{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(t)\, dt, \qquad x \in \mathbb{R}, \tag{1.4}$$

and note that this is of the form (1.2) with $G_n(x) = \int_{-\infty}^{x} g_n(t)\, dt$, where $g_n(t) = \alpha_n^{-1} g(t\alpha_n^{-1})$. Furthermore, if $\alpha_n \to 0$ as $n \to \infty$, then it is easy to check that $G_n \to^w u$.

In recent years, interest has increasingly been focused on the asymptotic properties of $\hat{F}_n$ and its use in estimation theory. The a.s. (almost sure) convergence of $\hat{F}_n$ to $F$ was proved by Nadaraya [4], Winter [13] and Yamato [15], while Watson and Leadbetter [11] obtained the asymptotic normality of $\hat{F}_n$. Recently, Winter [14] showed that $\hat{F}_n$ has the Chung–Smirnov property, i.e., with probability one

$$\limsup_{n \to \infty} (2n/\log\log n)^{1/2} \sup_x |\hat{F}_n(x) - F(x)| \leqslant 1.$$

To understand better the utility of the integral type estimators $\hat{F}_n$, it is worthwhile to study the asymptotic behavior of the distribution of $\hat{F}_n$ (defined by (1.4)) evaluated at a random point having the structure of a $U$-statistic. Specifically, let $h(x_1, \ldots, x_m)$, symmetric in its $m$ arguments, be a Borel-measurable kernel (of degree $m$). Based on a random sample

$X_1,..., X_n$ $(n \geq m)$ from a population with cdf $F$, the corresponding $U$-statistic $U_n$ for the estimation of the expected value $\xi = Eh(X_1,..., X_m)$ is formed by averaging the kernel $h$ symmetrically over the observations, namely

$$U_n = \binom{n}{m}^{-1} \sum_{C_{n,m}} h(X_{i_1},..., X_{i_m}),$$

where $C_{n,m}$ denotes the set of all the $\binom{n}{m}$ combinations of $m$ distinct elements $\{i_1,..., i_m\}$ from $\{1,..., n\}$.

If $\bar{X}$ denotes the sample mean, it is well known (see, e.g., Ghosh [2]) that under certain regularity conditions, $F_n(\bar{X})$, properly normalized, has normal distribution, in the limit (see also, Ralescu and Puri [6], for the corresponding problems concerning the rate of convergence of $F_n(\bar{X})$ to normality, a law of iterated logarithm, and an invariance principle for $F_n(\bar{X})$). The present note is devoted to investigating conditions for the asymptotic normality of the statistic $\hat{F}_n(U_n)$. Such a statistic is useful in estimating a functional $\theta = F(\xi)$ if $F$ is unknown. Also, if $U_n$ is the sample mean $\bar{X}$ (in which case $\xi = E(X_1)$), one may use $\hat{F}_n(\bar{X})$ for testing the hypothesis that a smooth $F$ is symmetric about an unknown location $\xi$ against certain alternatives.

## 2. Main Theorem

Let $g$ be a probability density function on $\mathbb{R}$, and let $\{\alpha_n\}$ (bandwidths) be a sequence of positive real numbers such that $\alpha_n \to 0$ as $n \to \infty$. If $g_n(t) = \alpha_n^{-1} g(t\alpha_n^{-1})$ and $G_n(x) = \int_{-\infty}^{x} g_n(t)\, dt$, consider the perturbed empirical distribution function $\hat{F}_n$ defined by (1.2) corresponding to the sequence $\{G_n\}$.

Let

$$h^*(t) = E_F h(t, X_2,..., X_m) \tag{2.1}$$

and

$$\zeta = \mathrm{Var}_F[h^*(X_1)]. \tag{2.2}$$

Then, our main result, dealing with the central limit theorem for a normalized $\hat{F}_n(U_n)$, is as follows:

Theorem. *Assume that*

(i) $E_F h^2(X_1,..., X_m) < \infty$ *and* $\zeta > 0$,

(ii) $\int_{-\infty}^{\infty} |t|\, g(t)\, dt < \infty$, *and*

(iii) $F$ *is twice differentiable on* $\mathbb{R}$ *with a bounded second derivative* $F''$.

*Then*

$$n^{1/2}[\hat{F}_n(U_n) - \mu_n] \xrightarrow{\mathcal{D}} N(0, \sigma^2), \tag{2.3}$$

*where* $\mu_n = EG_n(\xi - X_1) = \int_{-\infty}^{\infty} F(\xi - t)\, g_n(t)\, dt$ *and*

$$\sigma^2 = \mathrm{Var}[u(\xi - X_1) + m(h^*(X_1) - \xi)\, F'(\xi)] > 0.$$

*Proof.* Letting

$$Q_n = n^{1/2}[\hat{F}_n(U_n) - \hat{F}_n(\xi) - (U_n - \xi)\, F'(\xi)],$$

we first show that

$$Q_n \xrightarrow{P} 0. \tag{2.4}$$

To this end, set

$$Q_n = A_n + B_n,$$

where

$$A_n = \int [Z_n(U_n - t) - Z_n(\xi - t)]\, dG_n(t)$$

and

$$B_n = \sqrt{n} \int \{F(U_n - t) - F(\xi - t) - F'(\xi)(U_n - \xi)\}\, dG_n(t).$$

Here $Z_n = \sqrt{n}[F_n - F]$ is the standard empirical process, and $G_n$ is the df associated with the kernel in constructing $\hat{F}_n$.

Since $U_n - \xi = O(a_n)$ a.s., where $a_n = (n_{-1} \log\log n)^{1/2}$ under our assumptions (see [9, p. 191]), using the fluctuation result of Stute [10], we have that

$$|A_n| \leqslant \sup |Z_n(x) - Z_n(y)| \cdot \int dG_n(t)$$

(where the sup is taken over all $x$ and $y$, with $|x - y| \leqslant c a_n$)

$$= O\left( a_n \log \frac{1}{a_n} \right)^{1/2} \quad \text{a.s.}$$

for some sufficiently large $c > 0$ as $n \to \infty$. In the second term $B_n$, by using

the Lagrange form of Taylor expansion with remainder of $F(U_n - t)$ w.r.t. $\xi - t$ up to the second derivative term, the integrand equals

$$[F'(\xi - t) - F'(\xi)][U_n - \xi] + F''(\alpha(t))[U_n - \xi]^2/2$$
$$= F''(\beta(t))[U_n - \xi](-t) + F''(\alpha(t))[U_n - \xi]^2/2$$

with $|\beta(t) - \xi| \leqslant |t|$ and $|\alpha(t) - \xi + t| \leqslant |U_n - \xi|$ a.s.

Hence

$$|B_n| \leqslant \sqrt{n}\,|U_n - \xi| \cdot \|F''\|_\infty \cdot \int |t|\, dG_n(t)$$

$$+ \sqrt{n}\,|U_n - \xi|^2 \cdot \|F''\|_\infty \cdot \int dG_n(t)/2,$$

where $\|F''\|_\infty = \sup_x |F''(x)|$. Noting that $n^{1/2}(U_n - \xi) \to \mathcal{N}(0, m^2\zeta)$ in distribution, the first term on the right $\to 0$ in probability because of the choice of the bandwidth $\alpha_n$:

$$\int |t|\, dG_n(t) = \alpha_n \int |t|\, g(t)\, dt \to 0 \qquad \text{as} \quad n \to \infty.$$

The second term on the right $\to 0$ a.s. by using the Serfling result quoted earlier. Equation (2.4) follows.

Consequently, we can write

$$\hat{F}_n(U_n) = \hat{F}_n(\xi) + (U_n - \xi)\, F'(\xi) + n^{-1/2} Q_n. \tag{2.5}$$

Now, using condition (i) and Corollary 4.2 of Geertsema [1], we obtain the decomposition

$$U_n = \xi + \frac{m}{n} \sum_{i=1}^{n} \bar{h}(X_i) + R_n, \tag{2.6}$$

where $\bar{h}(t) = h^*(t) - \xi$, and for any $\rho > \frac{1}{2}$

$$R_n = o(n^{-1}(\log n)^\rho) \qquad \text{a.s. as } n \to \infty. \tag{2.7}$$

Therefore, from (2.5) and (2.6), we get the asymptotic representation

$$\hat{F}_n(U_n) = n^{-1} \sum_{j=1}^{n} T_{n,j} + F'(\xi)\, R_n + n^{-1/2} Q_n, \tag{2.8}$$

where $T_{n,j} = G_n(\xi - X_j) + mF'(\xi)\, \bar{h}(X_j)$, $1 \leqslant j \leqslant n$, $n \geqslant 1$. Since $G_n \to^w u$ entails $\mathrm{Var}(T_{n,1}) \to \sigma^2$ as $n \to \infty$, by using (2.4), (2.7) and (2.8) we see that the theorem will follow from the Lindeberg central limit theorem if we

show that the double-array $\{T_{n,j},\ 1 \leqslant j \leqslant n,\ n \geqslant 1\}$ of row-rise independent rv's satisfies the Lindeberg condition.

To this end, let $s_n^2 = \mathrm{Var}(\sum_{j=1}^n T_{n,j}) = n\,\mathrm{Var}(T_{n,1})$. For arbitary $\varepsilon > 0$, denote

$$L_n(\varepsilon) = s_n^{-2} \sum_{j=1}^{n} E\{(T_{n,j} - \mu_n)^2 I_{\{|T_{n,j} \quad \mu_n| > \varepsilon s_n\}}\}.$$

Now, using the inequality

$$(T_{n,j} - \mu_n)^2 \leqslant 2 + 2m^2[F'(\xi)]^2[\bar{h}(X_j)]^2,$$

we deduce that $L_n(\varepsilon) \leqslant L_{n,1}(\varepsilon) + L_{n,2}(\varepsilon)$, where

$$L_{n,1}(\varepsilon) = 2s_n^{-2} \sum_{j=1}^{n} P\{|T_{n,j} - \mu_n| > \varepsilon s_n\}$$

and

$$L_{n,2}(\varepsilon) = 2m^2[F'(\xi)]^2 s_n^{-2} \sum_{j=1}^{n} E\{[\bar{h}(X_j)]^2 I_{\{|T_{n,j} - \mu_n| > \varepsilon s_n\}}\}.$$

Now, since $s_n^2 \to \infty$ as $n \to \infty$, by Chebyshev's inequality it is easily seen that $L_{n,1}(\varepsilon) \to 0$ as $n \to \infty$.

To estimate $L_{n,2}(\varepsilon)$, note that for sufficiently large $n$,

$$\{|T_{n,j} - \mu_n| > \varepsilon s_n\} \subset \{mF'(\xi)|\bar{h}(X_j)| > \varepsilon s_n/2\}$$

which implies that

$$L_{n,2}(\varepsilon) \leqslant 2m^2[F'(\xi)]^2(\mathrm{Var}(T_{n,1}))^{-1} \int_A [\bar{h}(x)]^2\, dF(x) \tag{2.9}$$

where $A = \{2mF'(\xi)|\bar{h}(x)| > \varepsilon s_n\}$.

Then, since $\mathrm{Var}(T_{n,1}) \to \sigma^2$ as $n \to \infty$ and $\int_{-\infty}^{\infty} [\bar{h}(x)]^2\, dF(x) < \infty$, we see from (2.9) that $L_{n,2}(\varepsilon) \to 0$ as $n \to \infty$. Consequently, the conclusion $L_n(\varepsilon) \to 0$ as $n \to \infty$ obtains and the proof is completed.

EXAMPLE 1. The sample mean $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, for which $h(x) = x$, clearly satisfies condition (i) of our theorem if $0 < \mathrm{Var}(X_1) < \infty$.

EXAMPLE 2. The sample variance $S^2 = (n-1)^{-1}(\sum_{i=1}^n X_i^2 - n\bar{X}^2)$ corresponding to the kernel $h(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2 - 2x_1x_2)$ will satisfy condition (i) of our theorem provided that

$$0 < [\mathrm{Var}(X_1)]^2 < E\{(X_1 - E(X_1))^4\} < \infty.$$

EXAMPLE 3. For the Wilcoxon one-sample statistic $(h(x_1, x_2) = I_{\{x_1+x_2\leqslant 0\}})$, it is easily seen that if $0 < P_F\{X_1 + X_2 \leqslant 0\} < 1$, $\zeta > 0$ so that condition (i) of our theorem is automatically satisfied.

## ACKNOWLEDGMENT

The authors would like to express their sincere thanks to the referee for the very careful examination of the paper and for providing a shorter and elegant proof of the assertion (2.4).

## REFERENCES

1. GEERTSEMA, J. C. (1970). Sequential confidence intervals based on rank tests. *Ann. Math. Statist.* **41** 1016–1026.
2. GHOSH, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42** 1957–1961.
3. HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.
4. NADARAYA, E. A. (1964). Some new estimates for distribution functions. *Theory Probab. Appl.* **9** 497–500.
5. PARZEN, E. (1962). On the estimation of a probability density and mode. *Ann. Math. Statist.* **33** 1065–1076.
6. RALESCU, S. S., AND PURI, M. L. (1983). On Berry–Esséen rates, a law of the iterated logarithm and an invariance principle for the proportion of the sample below the sample mean. *J. Multivariate Anal.* **14** 231–247.
7. ROSENBLATT, M. (1956). Remark on some non-parametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.
8. SCOTT, D. W., TAPIA, R. A., AND THOMPSON, J. R. (1977). Kernel density estimation revisited. *Nonlinear Anal. Theor. Math. Appl.* **1** 339–372.
9. SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.
10. STUTE, W. (1982). The oscillation behavior of empirical processes. *Ann. Probab.* **10** 86–107.
11. WATSON, G. S., AND LEADBETTER, M. R. (1964). Hazard analysis II. *Sankhyā Ser. A.* **26** 101–116.
12. WERTZ, W. (1978). *Statistical Density Estimation: A Survey.* Vandenhoeck & Ruprecht, Gottingen.
13. WINTER, B. B. (1973). Strong uniform consistency of integrals of density estimators. *Canad. J. Statist.* **1** 247–253.
14. WINTER, B. B. (1979). Convergence rate of perturbed empirical distribution functions. *J. Appl. Probab.* **16** 163–173.
15. YAMATO, H. (1973). Uniform convergence of an estimator of a distribution function. *Bull. Math. Statist.* **15** 69–78.